JOHNS HOPKINS UNIVERSITY
MLDM - Final Project Report

An adapted sLDA on heart condition classification using PCG signal

Iskandar Atahodjaev, Aritro Mukherjee, Christos Sapsanis

May 18, 2017

1 Motivation

The main scope for this project is to determine if a single heart beat waveform corresponds to a healthy or an unhealthy subject. In our case, we have selected to use a phonocardiogram signal (PCG). The motivation relies on the fact that PCG is a common method for fast screening. Despite that fact, it is not as accurate for the specificity of a disease as the Echocardiogram (ultrasound), which is time consuming and needs specialized human resources. Thus, considering the time and human resources constrains, the main task is to achieve high accuracy as for the heart condition of the patient. In this work, a supervised version of Latent Dirichlet allocation (LDA) [1], sLDA [2], will be evaluated as a classification technique compared to Hidden Markov models [3], which is are commonly used for PCG signal classification [4].

2 Database

The database consists of samples collected from normal and abnormal subjects with variable heart condition. In the abnormal part, there are examples of Aortic Stenosis, Atrial septal defect, bicuspid aortic valve, double outlet right ventricle, Pulmonary stenosis, ventricular septal defect. Moreover, each of the recordings are 20 sec, which means that multiple heart sounds can be detected from each file. The number can vary due to the heart condition of the subject from 15 to 30 pulses. Moreover, for each subject, there are non-simultaneous measurements from up to five different positions around the heart. For each measurement, except from the PCG signal, there is also a EKG signal simultaneously measured, which will mainly be used for segmenting the PCG recording into dinstict heart beats. Given the above, we selected a subset, which included 36 labeled as healthy subjects and 41 labeled as unhealthy (mainly aortic stenosis).

3 Materials and Methods

The general procedure is described in Figure 3.1 and explained in details in this section. The main parts are three: signal pre-processing of the recording (segmentation, filtering and normalization), generation of feature collection (feature extraction and assigning a word to each feature) and classification (sLDA versus HMM).

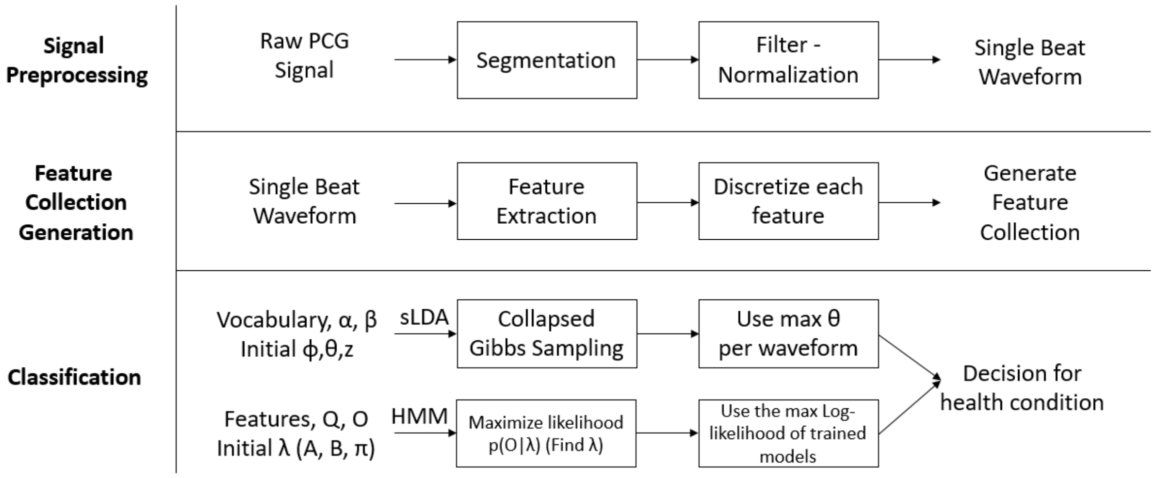


Figure 3.1: General Procedure.

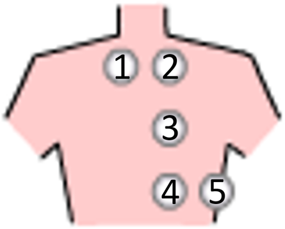


Figure 3.2: Locations of the measurements.

3.1 Signal Pre-processing

The position 4, as illustrated in Figure 3.2, recording were selected since it was the location that provided a more distinct separation between healthy and unhealthy subjects. The database used for this work included 797 healthy pulses and 934 unhealthy pulses (mainly aortic stenosis). The segmentation of the main recording used the simultaneous electrocardiogram (ECG), which eased the process. Afterwards, each segmented waveform was filtered using a 4^{th} order Butterworth Bandbass filter (20Hz - 250Hz), since this is the spectrum range under focus for PCG signal, and a 2^{nd} order IIR notch (60Hz) to remove the power supply interference. The last step was to normalize the full signal between values -1 and +1.

3.2 Feature Extraction

The two main categories for feature extraction that has be used are the envelogram and the Mel Frequency Cepstral Coefficients (MFCCs), as depicted in Figure 3.3 and Figure 3.4 respectively. They commonly used techniques in literature and, thus, no more analysis will be conducted in the current report.

3.3 Mapping signal's values into words.

An important aspect of this work is the representation of the signal value into words. The procedure that was followed is depicted in Figure 3.5. Each time point (in overall 100 points per signal) for each signal was used to construct a histogram. Each bin corresponds to a word-type in the vocabulary. Hence, each value in a specific bin will be assigned with the corresponding word-type. This word-type cannot be used by another time point, which had it's own distinct. Thus, the overall collection included the vocabularies for each time point, which preserves the information for time order. Another approach for mapping audio words and use LDA is presented in [5], where each segment of a spoken word was tried to be assigned to a word-type.

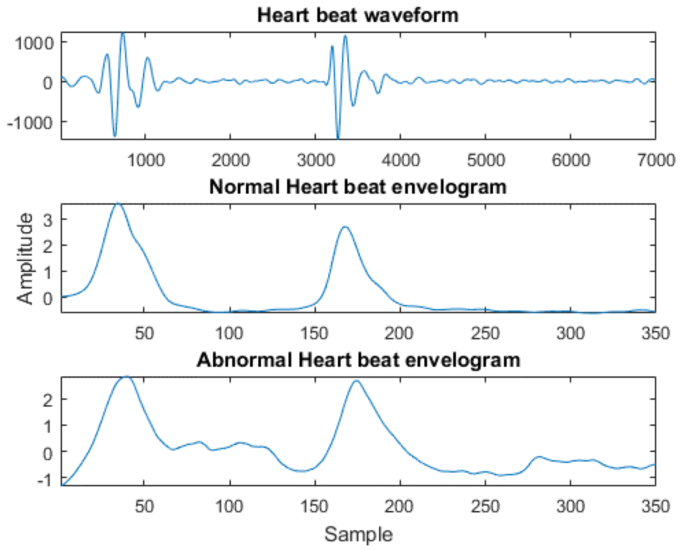


Figure 3.3: Envelogram.

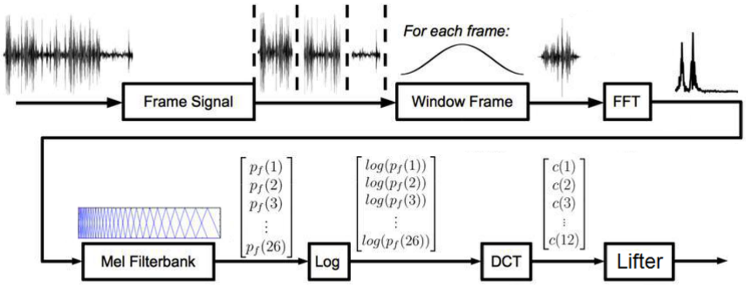

For each frame:
Frame Signal
Window Frame
FFT
Mel Filterbank
$p_f(1)$ $p_f(2)$ $p_f(3)$ ⋮ $p_f(26)$
Log
$log(p_f(1))$ $log(p_f(2))$ $log(p_f(3))$ ⋮ $log(p_f(26))$
DCT
$c(1)$ $c(2)$ $c(3)$ ⋮ $c(12)$
Lifter

Figure 3.4: Mel Frequency Cepstral Coefficients procedure.

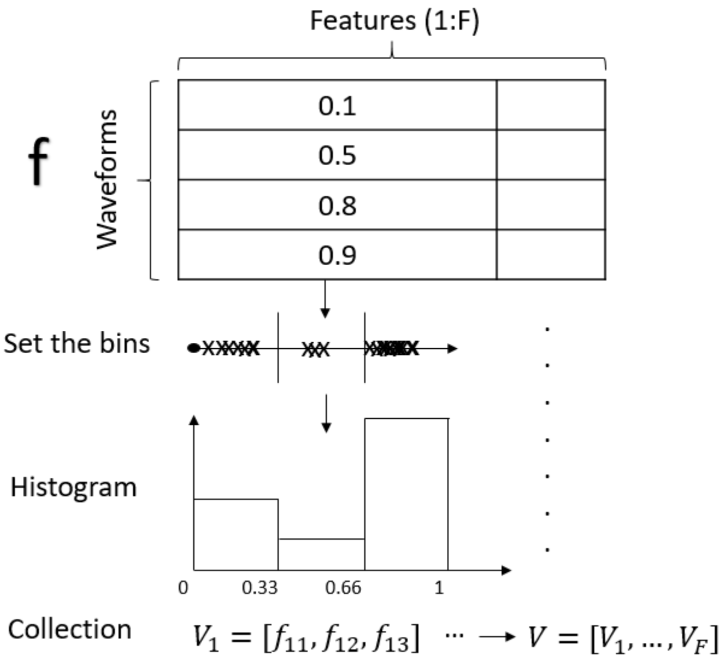
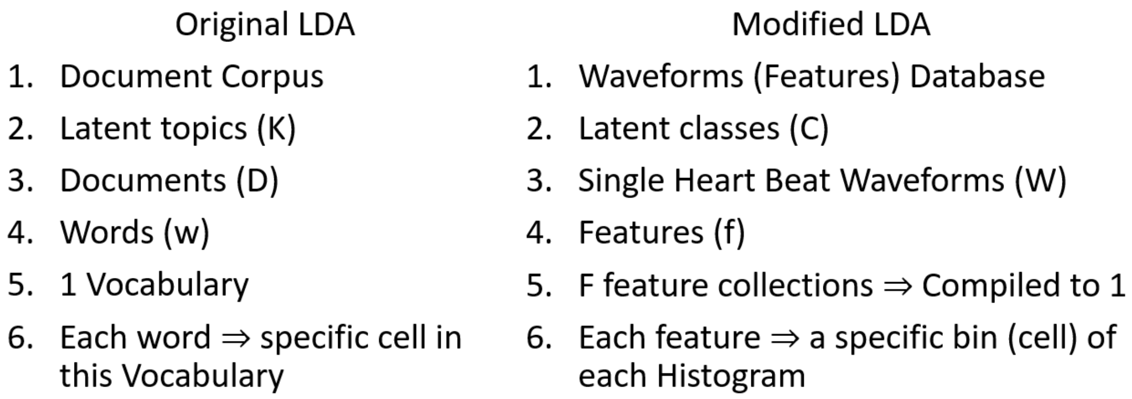


Figure 3.5: Mapping signal's values into words.



Original LDA	Modified LDA
1. Document Corpus	1. Waveforms (Features) Database
2. Latent topics (K)	2. Latent classes (C)
3. Documents (D)	3. Single Heart Beat Waveforms (W)
4. Words (w)	4. Features (f)
5. 1 Vocabulary	5. F feature collections ⇒ Compiled to 1
6. Each word ⇒ specific cell in this Vocabulary	6. Each feature ⇒ a specific bin (cell) of each Histogram

Figure 3.6: Correlation between original and modified LDA.

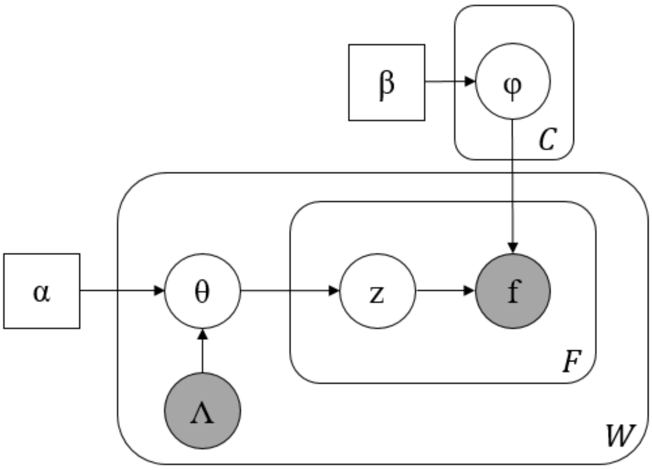


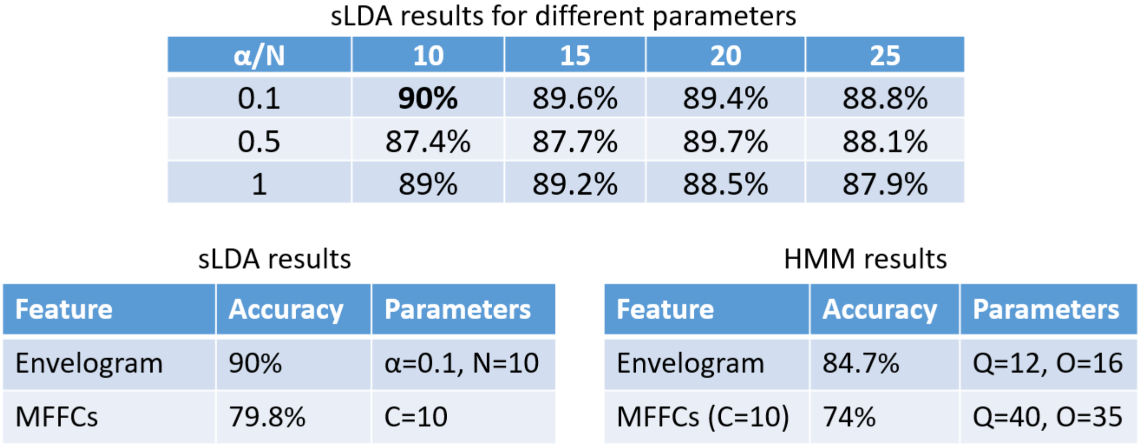
Figure 3.7: Supervised Latent Dirichlet Allocation.

3.4 Supervised Latent Dirichlet Allocation (sLDA)

After achieving to map the signal into words, LDA can be used. The reason for selecting LDA is that it has a nice analogy to our problem, as depicted in Figure 3.6. Each waveform can be a mixture of the classes θ - Dir(α), which is common for a patient to have more than one cardiovascular disease. Moreover, some feature values, as the words, can be similar/shared across different signals, as the documents (z - Mult(θ)). In our case, the labels for each signal are know so the task will be classification and not clustering, which is supported by LDA. So the initial LDA should be transformed into sLDA, specifically tailored for classification. In order to achieve this, a new observed random variable, Λ, is inserted to constrain the θ in the classes that this signal is assigned, as depicted in Figure 3.7.

4 Results Analysis

As it can be observed from Figure 4.1, the sLDA outperformed the HMM classification results reaching 90% compared to the 84.7% for using Envelogram. The main reason for this achievement is that sLDA took advantage of: LDA's correlation of topics with similar words, which is our case are the features, and it preserved the time sequence. As for every classification algorithm, the results are highly dependent on feature space and feature selection model and on the ability of the data to be separable. Thus, the use of different feature extraction approaches revealed some information about the data. The MFCCs, which include spectral information, except from the temporal, achieved a lower accuracy compared to Envelogram. The main reason for this is



sLDA results for different parameters

α/N	10	15	20	25
0.1	**90%**	89.6%	89.4%	88.8%
0.5	87.4%	87.7%	89.7%	88.1%
1	89%	89.2%	88.5%	87.9%

sLDA results

Feature	Accuracy	Parameters
Envelogram	90%	α=0.1, N=10
MFFCs	79.8%	C=10

HMM results

Feature	Accuracy	Parameters
Envelogram	84.7%	Q=12, O=16
MFFCs (C=10)	74%	Q=40, O=35

Figure 4.1: Classification Results for sLDA and HMM.

that the way that frequency information was encoded by MFCCs has not add anything that could separate the classes. On the contrary, it made them less separable. Thus, a more informative representation for each time segment MFCCs should be conducted in order to represent a word with more uniqueness among the classes.The Envelogram has provided the time series signature of the signal, which proved more efficient in the current approach.

5 Future work

The future work will be focused on three different parts related with database variability, feature extraction and models for classification.

- The first is to expand the available heart beat waveforms in the abnormal case, so more diseases (Atrial septal defect, Pulmonary stenosis, Ventricular septal defect etc.) will be represented in the abnormal case enhancing the variability of the dataset.
- The second part is to extract more features. The classification performance using MFCCs as a feature has not operated in the expected way. They can be re-estimated using a larger number of triangular filters in the spectrum area of interest (20-250Hz), in order to capture the features in a higher resolution. Moreover, features representing the bursts of energy will be extracted, such as the intensity, the duration (temporal width), the bandwidth (frequency range), the central frequency and the temporal center. Moreover, wavelet transform and empirical mode decomposition can also be utilized.
- The third part will be focused on the algorithms and models that can be used in a search of a more accurate classification. The first step will be to explore the topic modeling area, for example by using probabilistic latent semantic indexing (pLSI). Moreover, the area of Deep Neural Networks (DNNs) (multi-layer perceptron (MLP), Convolutional Neural Network (CNN)) will be examined with the an expansion of the database to avoid over-fitting. For instance, the MFCCs features, since they are a 2D array, can be used as an image and imported in a CNN. Moreover, a hand crafted tree augmented Naive Bayes classifier can be used, where the signal features will be nodes and can be dependence relationships in between them having all of them as a parent the class. In order to construct this network, a more deep understanding on the distinct differences between the classes will be needed.

In this project, the heart beat signal segmentation has been based on the EKG data that were recorded simultaneously. Another way to segment the signal will be to use HMM for detecting the states for the waveform.

6 Conclusion

In overall, to the best of our knowledge, this work was the first that used sLDA for phonocardiogram signal classification. The results for LDA seem promising since it outperformed the HMMs accuracy. Moreover, there are potentials for improving the results based on the future work plan.

7 Appendix (Code Explanation)

7.1 Feature Extraction

The feature extraction is conducted using PartB.m script. In the beginning, the signal is normalized from -1 to 1. Then, by using C. Potes' code from Physionet Challenge 2016, the envelogram is extracted using getSpringerPCGFeatures.m. Moreover, the MFFCs were computed by using the corresponding file (mfcc.m) and library provided by Mathworks. The full signal could be decomposed to 8 MFCCs per segment (55 in overall), which corresponded in the frequency range of 20-250 Hz. The envelogram had initially 400 points and it was down-sampled by to 100, since it could preserve the same information.

7.2 sLDA

The code for sLDA used collapsed Gibbs sampling. The first part conducts the discretization of the input and based on a histogram per time point, it corresponds to a specific word in the vocabulary. Since a supervised version of LDA was used, in the training part, the θ was constrained in this part giving values only to the class that correspond to the signal and setting the other one 0. As for the testing part, the rest 30% of the data was used for testing with a ϕ variable to be constant throughout all that period, as it was estimated in the previous step. After 150 iterations (burn in was 100), the θ values are averaged based on the 50 last iterations. Based on these values, the class for each test sample is assigned to one with higher θ.

7.3 HMM

The Murphy's library was mainly used for HMM. The main code for envelogram is main-test1-env.m and for MFCCs is main-test1-mfcc-vec.m. The HMM parameters are initialized by using the mk-stochastic.m, which generates random matrices that each row adds to 1. Then the data are shuffled randomly and 70% of them are used for training. For the estimating the parameters of the HMM model, the function dhmm-em.m was utilized which find the ML/MAP parameters of an HMM with discrete outputs using EM. Each class has the same structure with each other but it is trained with the corresponding input and so eventually it will have different parameters. The next step is to insert each test sample as an input in both trained HMMs. The one which provides the highest value log-likelihood value will be the estimated class. The dhmm-logprob.m function computes the log-likelihood of the test sample using the parameters for the discrete HMMs that were estimated before.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] D. Ramage, et al., “Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora,” *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, 2009.

[3] L. R. Rabiner, A. Y. Ng, and M. I. Jordan, “A tutorial on hidden Markov models and selected applications in speech recognition,” *IEEE Proceedings*, vol. 77.2, pp. 257–286, 1989.

[4] S. Chauhan, et al, “A computer-aided MFCC-based HMM system for automatic auscultation,” *Computers in Biology and Medicine*, vol. 38.2, pp. 221–233, 2008.

[5] S. Kim, P. Georgiou, and S. Narayanan, “Supervised acoustic topic model with a consequent classifier for unstructured audio classification,” *2012 10th International IEEE Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012.